Stream Processing for Remote Collaborative Data Analysis

Scott Klasky[146], C. S. Chang[2], Jong Choi[1], Michael Churchill[2], Tahsin Kurc[51], Manish Parashar[3], Alex Sim[7], Matthew Wolf[14], John Wu[7]

[1] ORNL, [2] PPPL, [3] Rutgers, [4]GT, [5] SBU, [6]UTK, [7] LBNL

STREAM 2016

Tysons, VA

Next Generation DOE computing

System attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrades	
Name Planned Installation	Edison	TITAN	MIRA	Cori 2016	Summit 2017-2018	Theta 2016	Aurora 2018-2019
System peak (PF)	2.6	27	10	> 30	150	>8.5	180
Peak Power (MW)	2	9	4.8	< 3.7	10	1.7	13
Total system memory	357 TB	710TB	768TB	~1 PB DDR4 + High Bandwidth Memory (HBM)+1.5PB persistent memory	> 1.74 PB DDR4 + HBM + 2.8 PB persistent memory	>480 TB DDR4 + High Bandwidth Memory (HBM)	> 7 PB High Bandwidth On-Package Memory Local Memory and Persistent Memory
Node performance (TF)	0.460	1.452	0.204	> 3	> 40	> 3	> 17 times Mira
Node processors	Intel Ivy Bridge	AMD Opteron Nvidia Kepler	64-bit PowerPC A2	Intel Knights Landing many core CPUs Intel Haswell CPU in data partition	Multiple IBM Power9 CPUs & multiple Nvidia Voltas GPUS	Intel Knights Landing Xeon Phi many core CPUs	Knights Hill Xeon Phi many core CPUs
System size (nodes)	5,600 nodes	18,688 nodes	49,152	9,300 nodes 1,900 nodes in data partition	~3,500 nodes	>2,500 nodes	>50,000 nodes
System Interconnect	Aries	Gemini	5D Torus	Aries	Dual Rail EDR-IB	Aries	2nd Generation Intel Omni-Path Architecture
File System	7.6 PB 168 GB/s,	32 PB 1 TB/s, Lustre®	26 PB 300 GB/s GPFS™	28 PB 744 GB/s Lustre®	120 PB 1 TB/s GPFS™	10PB, 210 GB/s Lustre initial	150 PB 1 TB/s Lustre®

- File System and network bandwidth does not keep up with computing power

Big Data in Fusion Science: ITER example

- **Volume**: Initially 90 TB per day, 18 PB per year, maturing to 2.2 PB per day, 440 PB per year
- **Value**: All data are taken from expensive instruments for valuable reasons.
- **Velocity**: Peak 50 GB/s, with near real-time analysis needs
- **Variety**: ~100 different types of instruments and sensors, numbering in the thousands, producing interdependent data in various formats
- **Veracity**: The quality of the data can vary greatly depending upon the instruments and sensors.

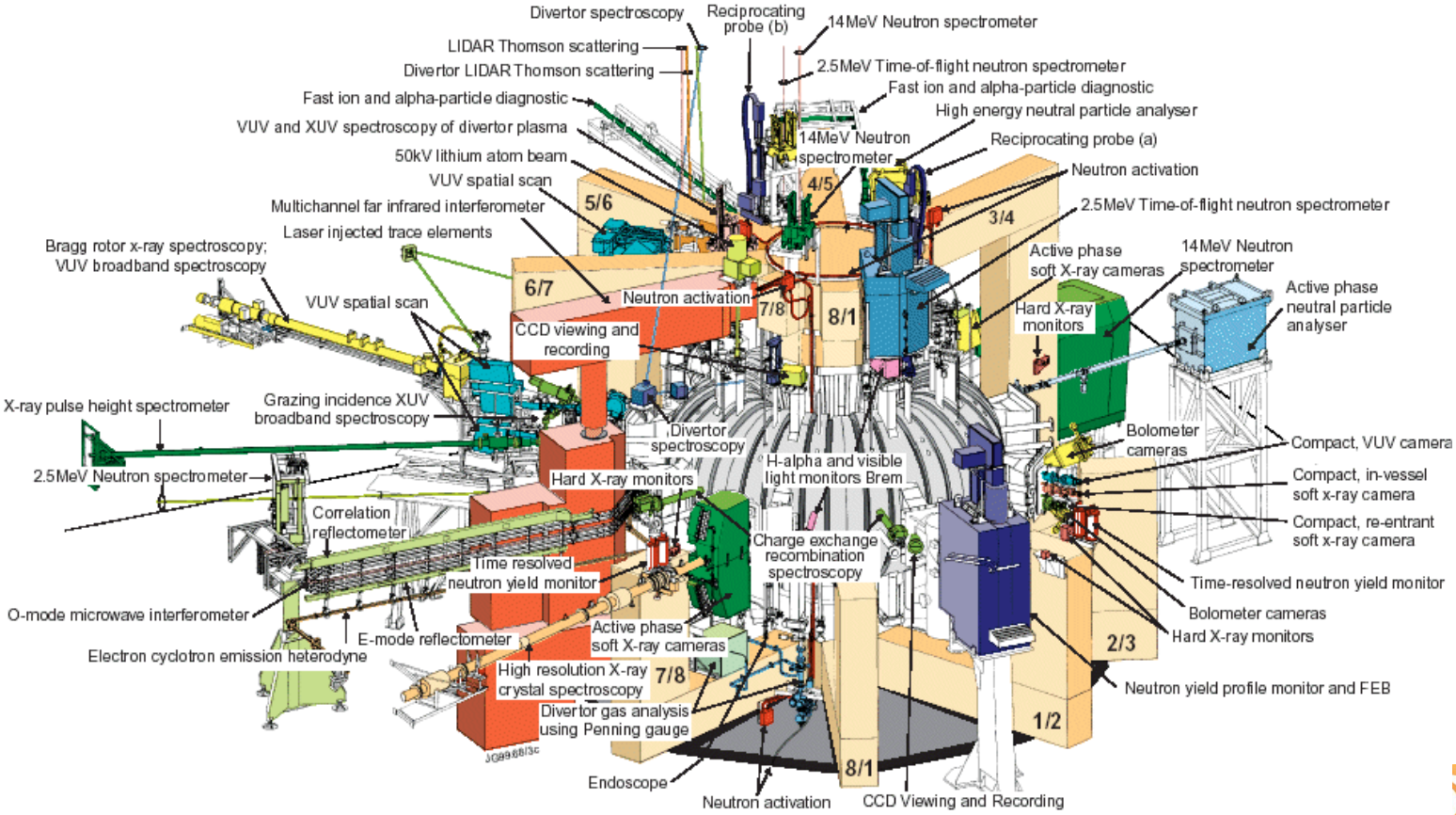
The pre-ITER superconducting fusion experiments outside of US will also produce increasingly bigger data (KSTAR, EAST, Wendelstein 7-X, and JT60-SU later).

Streaming data from Simulations

- Whole Device Modeling – the next fusion grand challenge
  - From 1992 NSF grand challenge
  - Now a possible Exascale Computing Application
  - "Much more difficult than original thought"
  - Couple Codes from different time scales, different physics, different spatial regimes.
- Very large data
  - Edge simulation "today" producing about 100 PB in 10 days.
  - Couple MHD effects, Core effects, …. → 1 EB in 2 days.
  - Need data triage techniques
  - Separate out information from data in near-real-time
- Very large number of observables
  - Large number of workflows for each set of analysis
- Desire to understand what do we do on machine, off machine

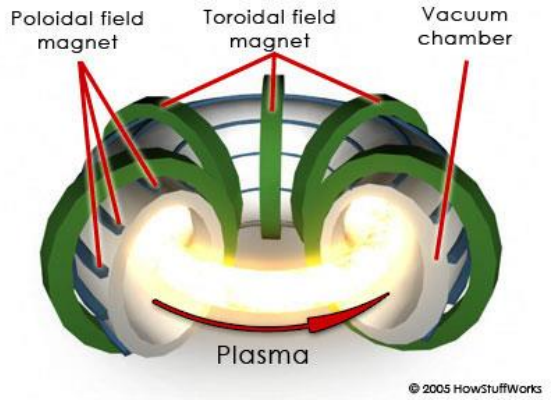
Collaborative Nature of Science

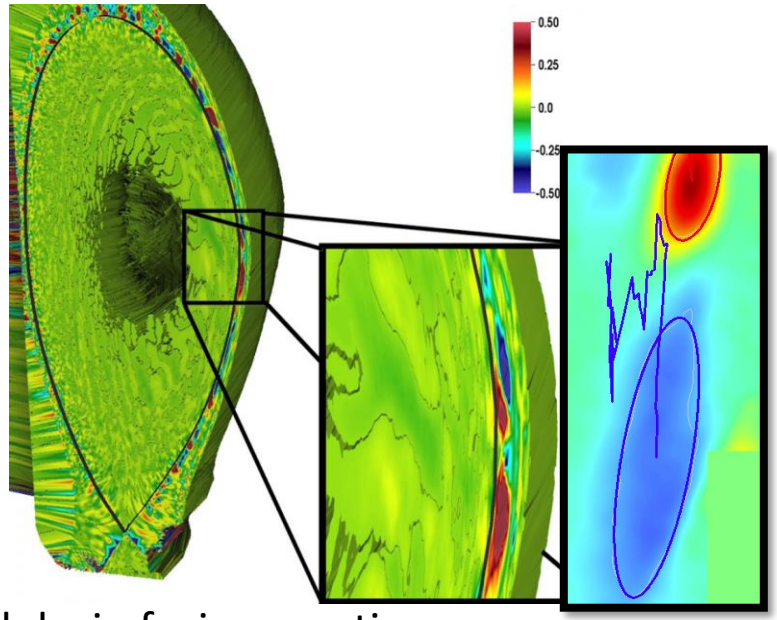
- 100's different diagnostics which produce data of different sizes/velocities
- Different scientist have workflows
- Fusion simulations produce 100's of different variables
- Goal: run all analysis to make near-real-time decisions
- Realization: V's are too large: Prioritize which data gets processed when, where



Feature Extraction: Near Real Time Detection of Blobs

- Fusion Plasma blobs
  - Lead to the loss of energy from tokamak plasmas
  - Could damage multi-billion tokamak
- The experimental facility may not have enough computing power for the necessary data processing
- Distributed in transient processing
  - Make more processing power available
  - Allow more scientists to participate in the data analysis operations and monitor the experiment remotely
  - Enable scientists to share knowledge and processes





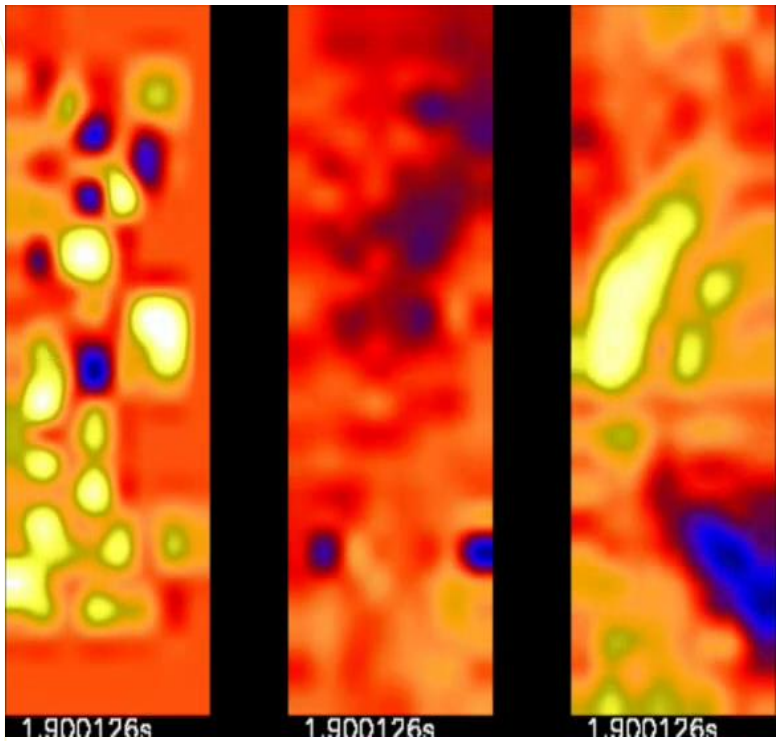
Blobs in fusion reaction (Source: EPSI project)

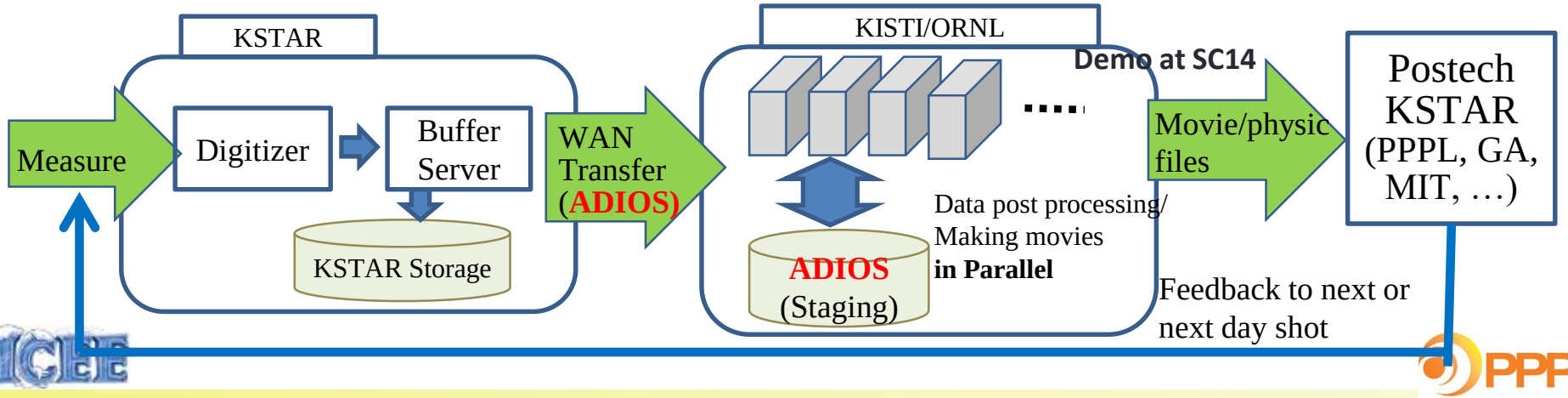
Blob trajectory

ICEE, Enabling International Collaborations

Example: KSTAR ECEI Sample Workflow

- **Objective**: To enable remote scientists to study ECE-Image movies of blobby turbulence and instabilities between experimental shots in near real-time.
- **Input**: Raw ECEi voltage data (~550MB/s, over 300 seconds in the future) + Metadata (experimental setting)
- **Requirement**: Data transfer, processing, and feedback **within <15min** (inter-shot time)
- **Implementation**: distributed data processing with ADIOS ICEE method



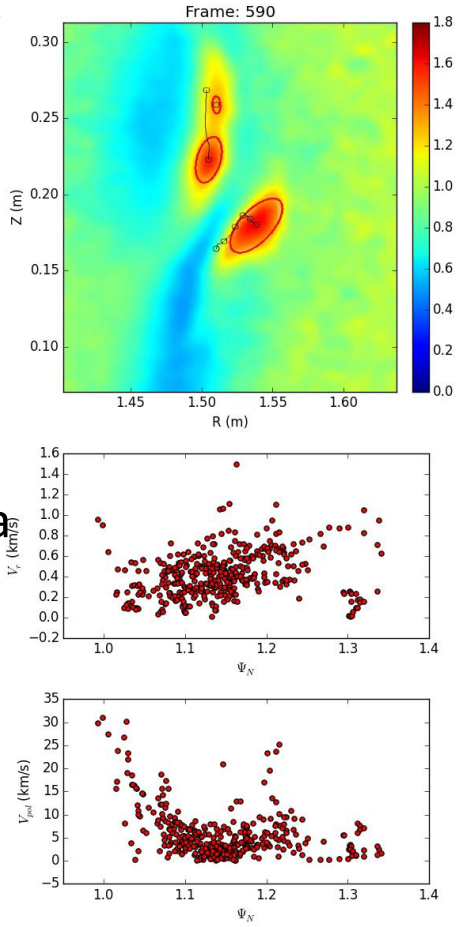


ICEE, Enabling Fusion Collaboration

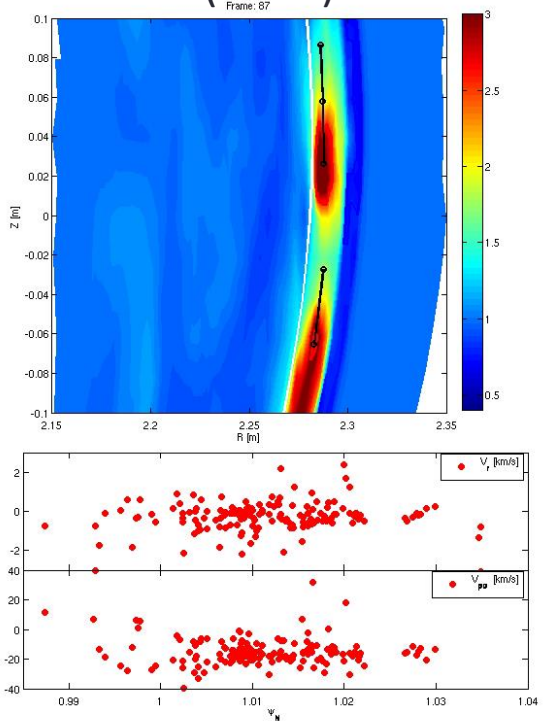
Data Fusion

- **Objective**: Enable comparisons of simulation (pre/post) and experiment at remote locations
- **Input**: Gas Puff Imaging (GPI) fast camera images from NSTX and XGC1 edge simulation data
- **Output**: Blob physics
- **Requirement**: Complete in near real-time for inter-shot experimental comparison, experiment-simulation validation or simulation monitoring
- **Implementation**: distributed data processing with ADIOS ICEE method, optimized detection algorithms for near real-time analysis

Experiment (NSTX GPI)



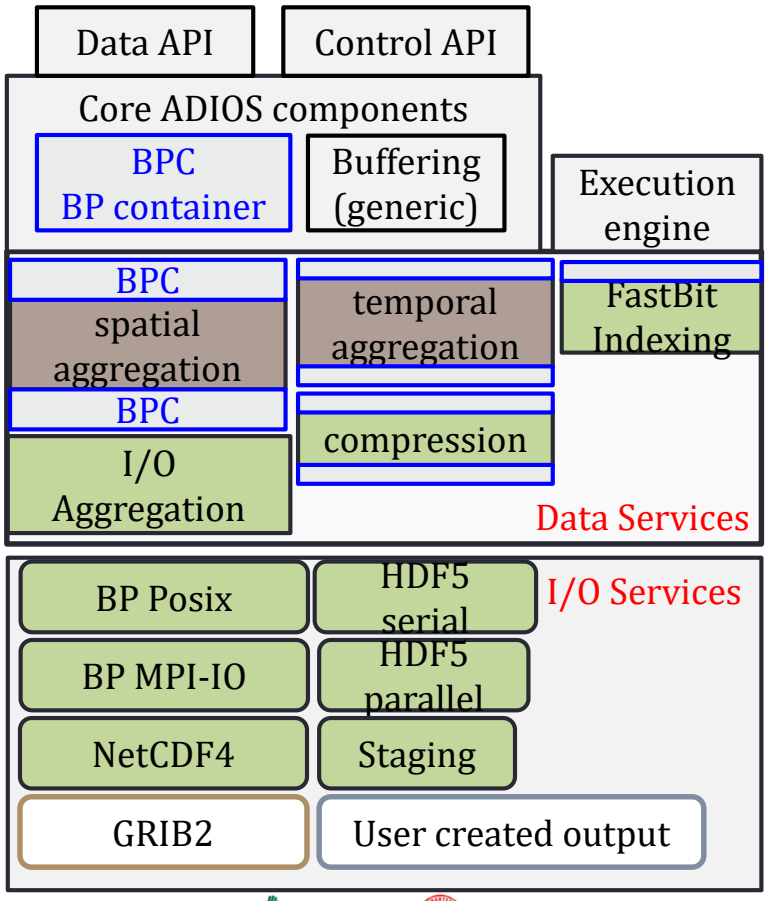
Simulation (XGC1)



ADIOS Abstraction Unifies Local And Remote I/O

- I/O Componentization for Data-at-Rest and Data-in-Motion
- Service Oriented Architecture for Extreme scaling computing
- Self Describing data movement/storage
- Main paper to cite

Q. Liu, J. Logan, Y. Tian, H. Abbasi, N. Podhorszki, J. Choi, S. Klasky, R. Tchoua, J. Lofstead, R. Oldfield, M. Parashar, N. Samatova, K. Schwan, A. Shoshani, M. Wolf, K. Wu, W. Yu, "Hello ADIOS: the challenges and lessons of developing leadership class I/O frameworks", Concurrency and Computation: Practice and Experience, 2013



ADIOS applications

1. Accelerator: PIConGPU, Warp
2. Astronomy: SKA
3. Astrophysics: Chimera
4. Combustion: S3D
5. CFD: FINE/Turbo, OpenFoam
6. Fusion: XGC, GTC, GTC-P, M3D,M3D-C1, M3D-K, Pixie3D
7. Geoscience: SPECFEM3D_GLOBE, AWP-ODC, RTM
8. Materials Science: QMCPack, LAMMPS
9. Medical Imaging: Cancer pathology
10. Quantum Turbulence: QLG2Q
11. Relativity: Maya
12. Weather: GRAPES
13. Visualization: Paraview, Visit, VTK, ITK, OpenCV, VTKm

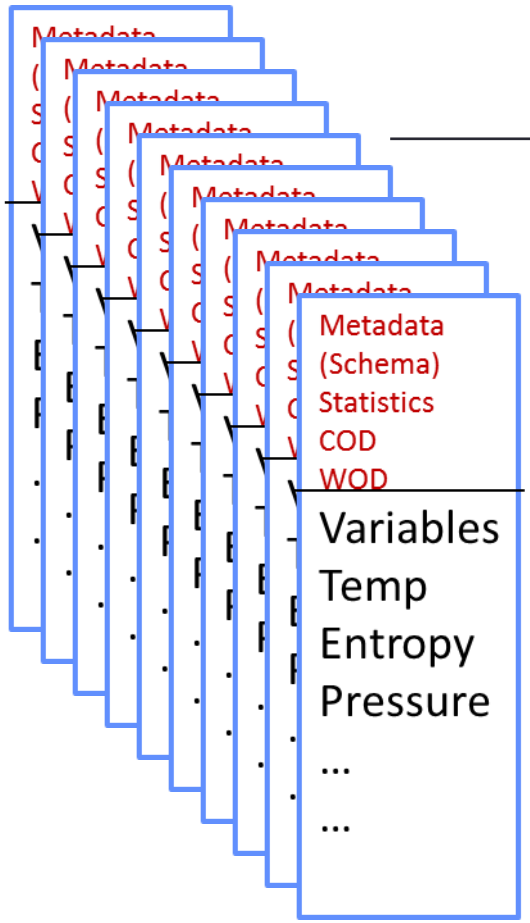
Impact on Industry :

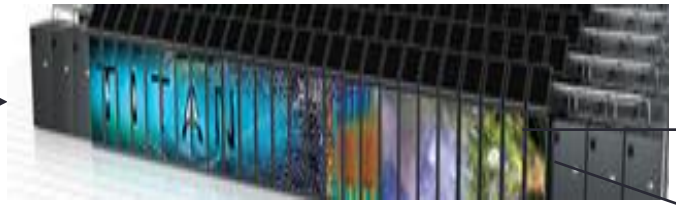
- NUMECA (FINE/Turbo) – Allowed time-varying interaction of turbomachinery-related aerodynamic phenomena
- TOTAL (RTM) – Allowed running of higher fidelity seismic simulations
- FMGLOBAL (OpenFoam)– Allowed running higher fidelity fire propagation simulations

Over 1B LCF hours from ADIOS enabled Apps 2015
Over 1,500 citations

LCF/NERSC Codes in red

The ADIOS-BP Stream/File format





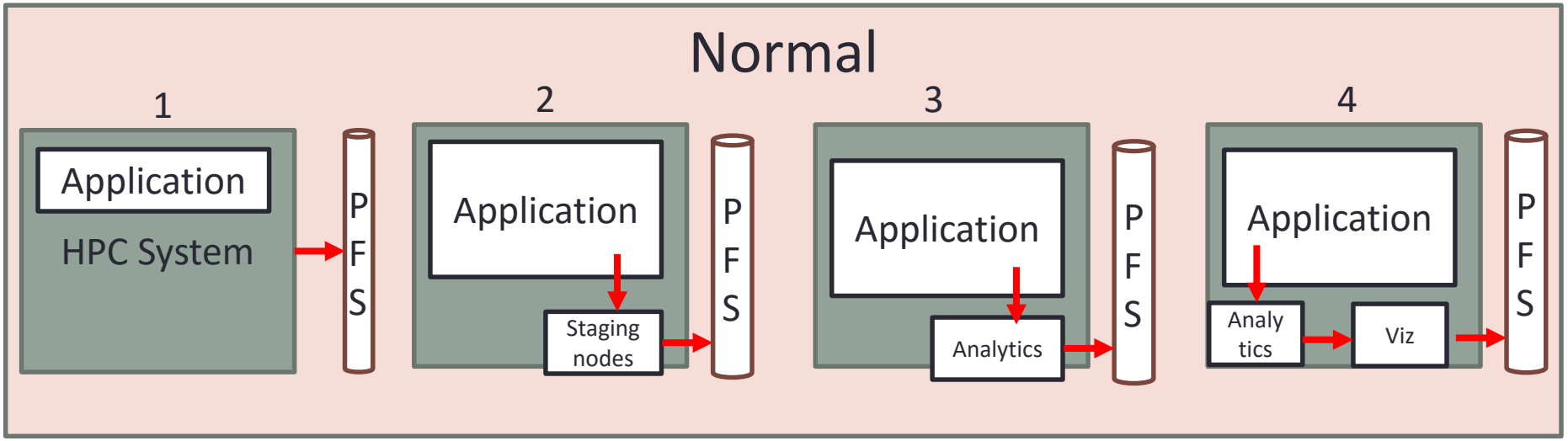
- All data chunks are from a single producer
  - MPI process, Single diagnostic
- Ability to create a separate metadata file when "sub-files" are generated
- Allows variables to be individually compressed
- Has a schema to introspect the information
- Has workflows embedded into the data streams
- Format is for "data-in-motion" and "data-at-rest"

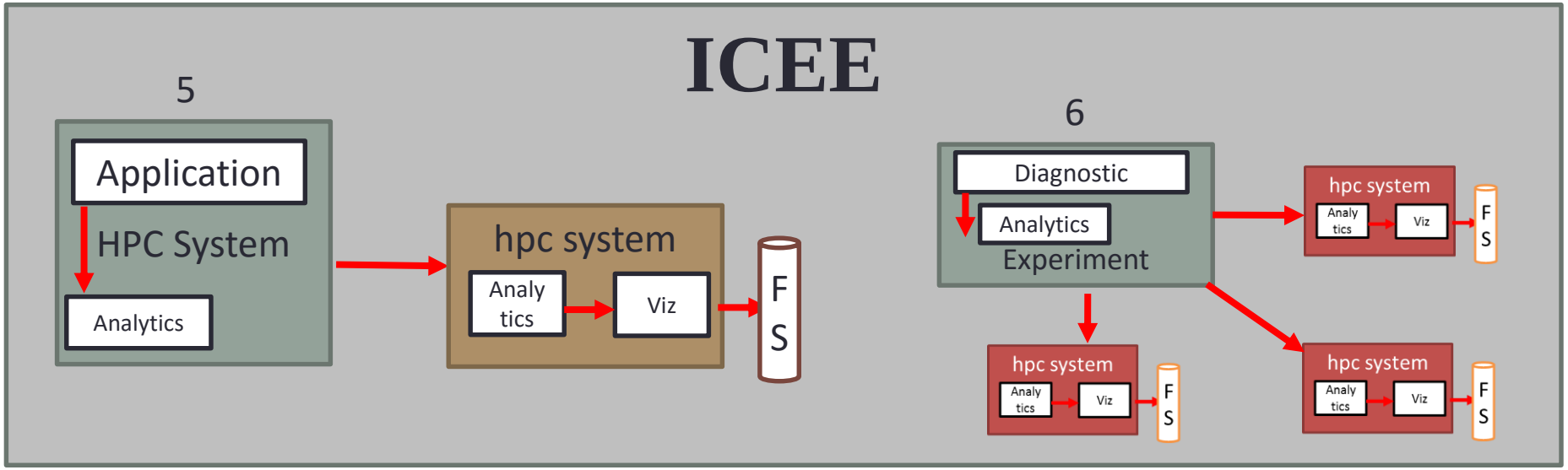
Ensemble of chunks = file

Auditing data during the streaming phase

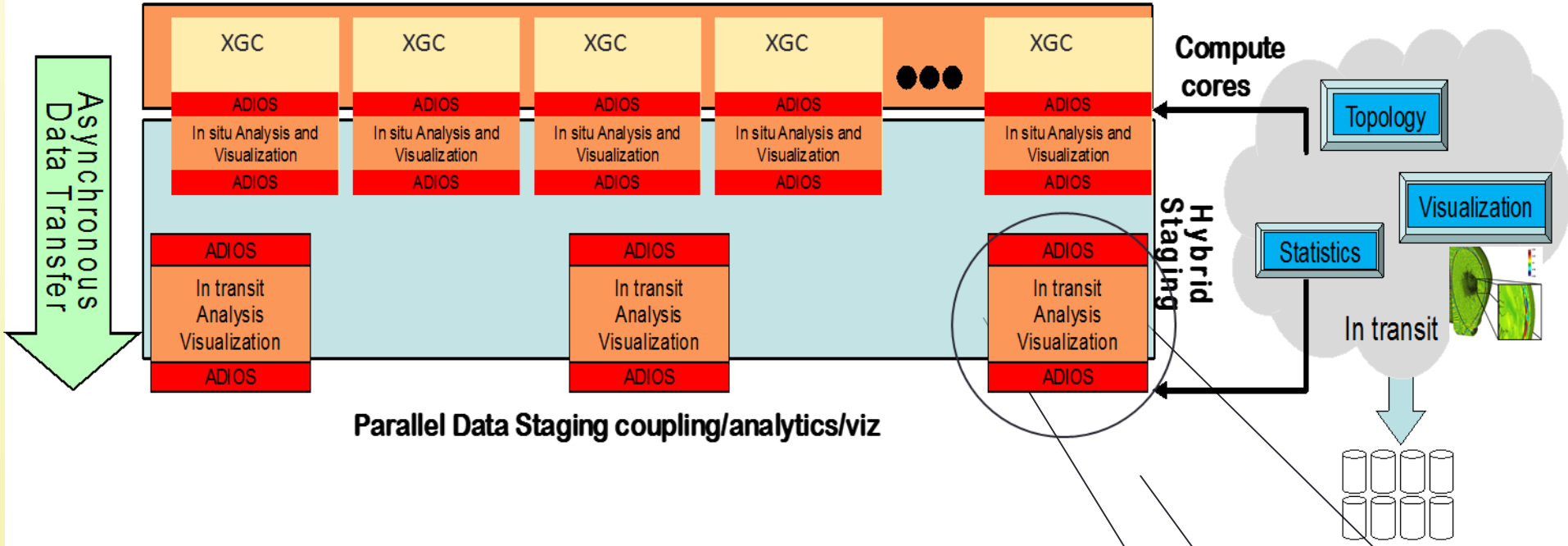
- ~~Streaming Data~~ **Stream Information**
  - Too much data to move in too little time
  - Storage sizes/speed doesn't keep up with making NRT decisions
- Fit the data with a model
  - From our physics understanding
  - From our understanding of the entropy in the data
  - From our understanding of the error in the data
  - Change data into data = model + information ($f = F + \delta f$)
- Streaming Information
  - Reconstruct "on-the-fly"
- Query data from the remote source

I/O abstraction of data staging



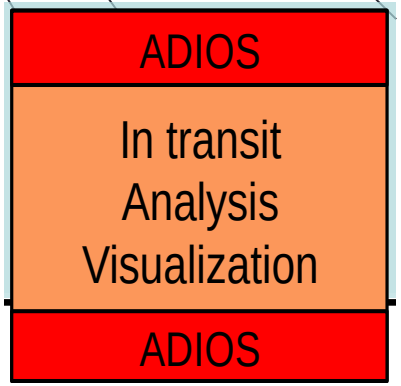


Hybrid Staging

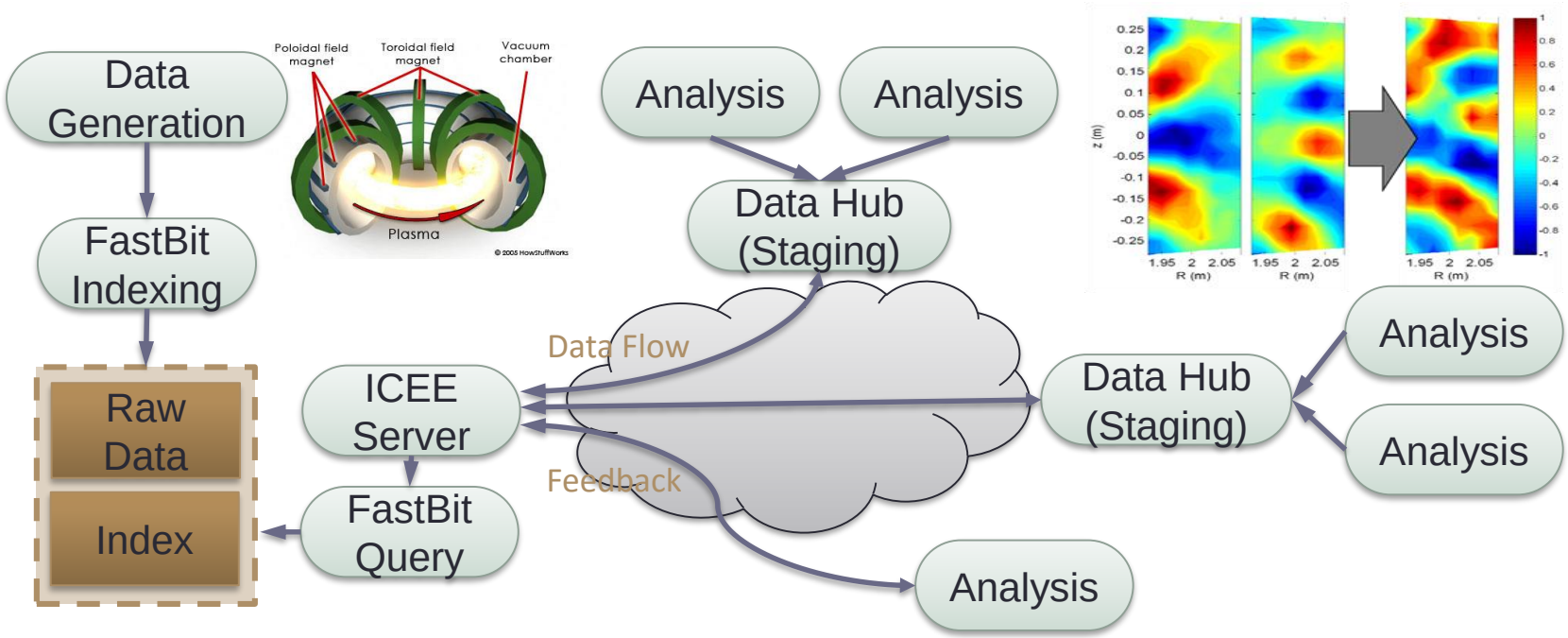


- Use compute and deep-memory hierarchies to optimize overall workflow for power vs. performance tradeoffs
- Abstract complex/deep memory hierarchy access
- Placement of analysis and visualization tasks in a complex system

- Impact of network data movement compared to memory movement
- Abstraction allows staging
  - On-same core
  - On different cores
  - On different nodes
  - On different machines
  - Through the storage system

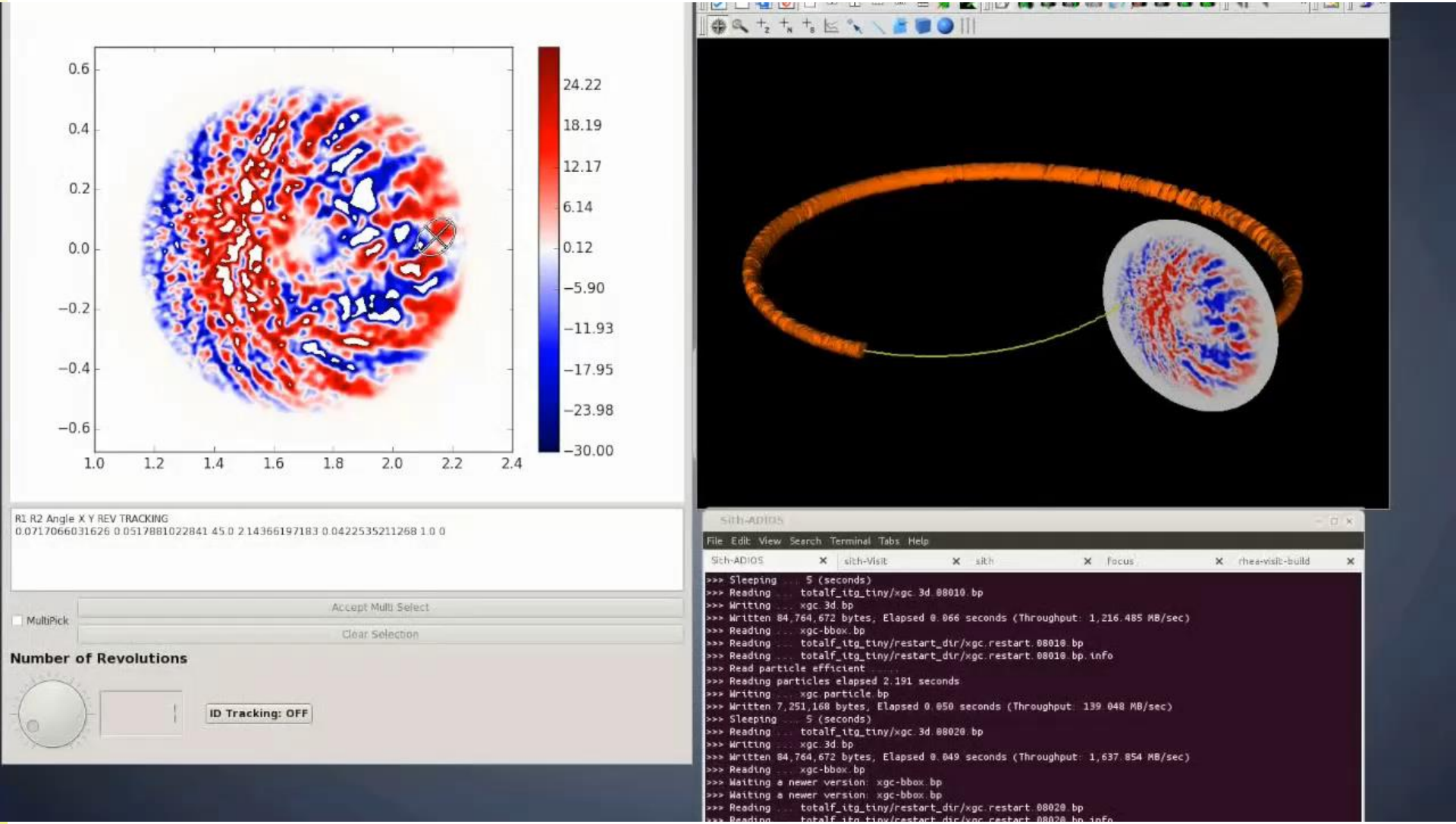


ICEE System Development With ADIOS



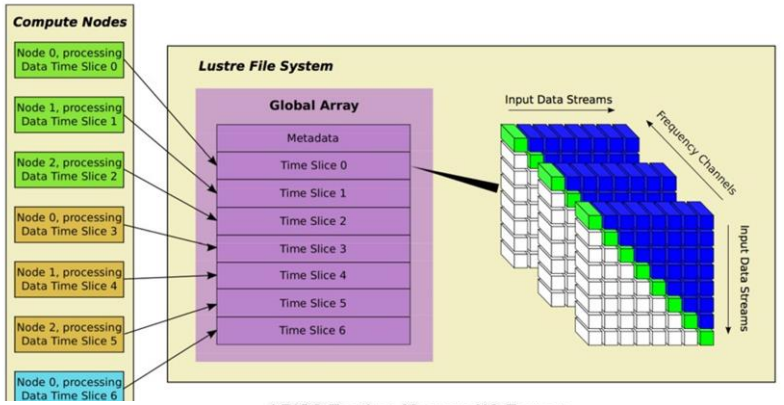
- Features
  - ADIOS provides an overlay network to share data and give feedbacks
  - Stream data processing – supports stream-based IO to process pulse data
  - In transit processing – provides remote memory-to-memory mapping between data source (data generator) and client (data consumer)
  - Indexing and querying with FastBit technology

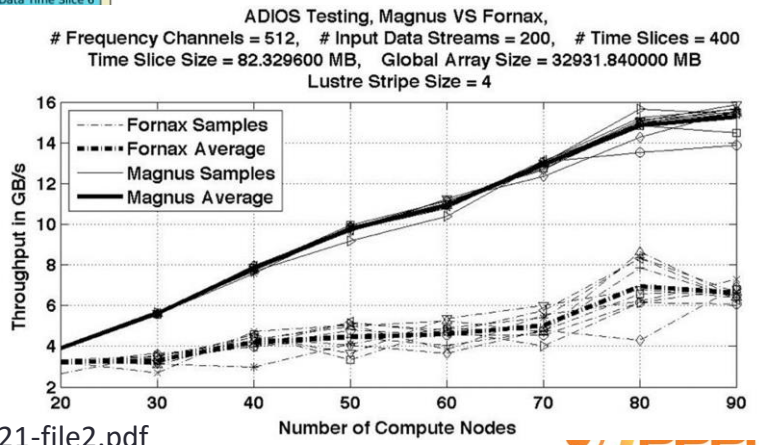
Interactive Supercomputing from Singapore to Austin



Technology is also being used for SKA (J. Wang)

- The largest radio telescope in the world
- €1.5 billion project
- 11 member countries
- **2023-2030** Phase 2 constructed
- Currently conceptual design & preliminary benchmarks !
- Compute Challenge: • 100 PFLOPS
- Data Challenge: ExaBytes per day
- Challenge is to run time-division correlator and then write output data to a parallel filesystem





https://cug.org/proceedings/cug2014_proceedings/includes/files/pap121-file2.pdf